Data Analysis and Visualization in R (IN2339)

Exercise Session 8 - Statistical Testing II

Hasan Celik, Christian Mertes, Vicente Yépez

Section 00 - Getting Ready

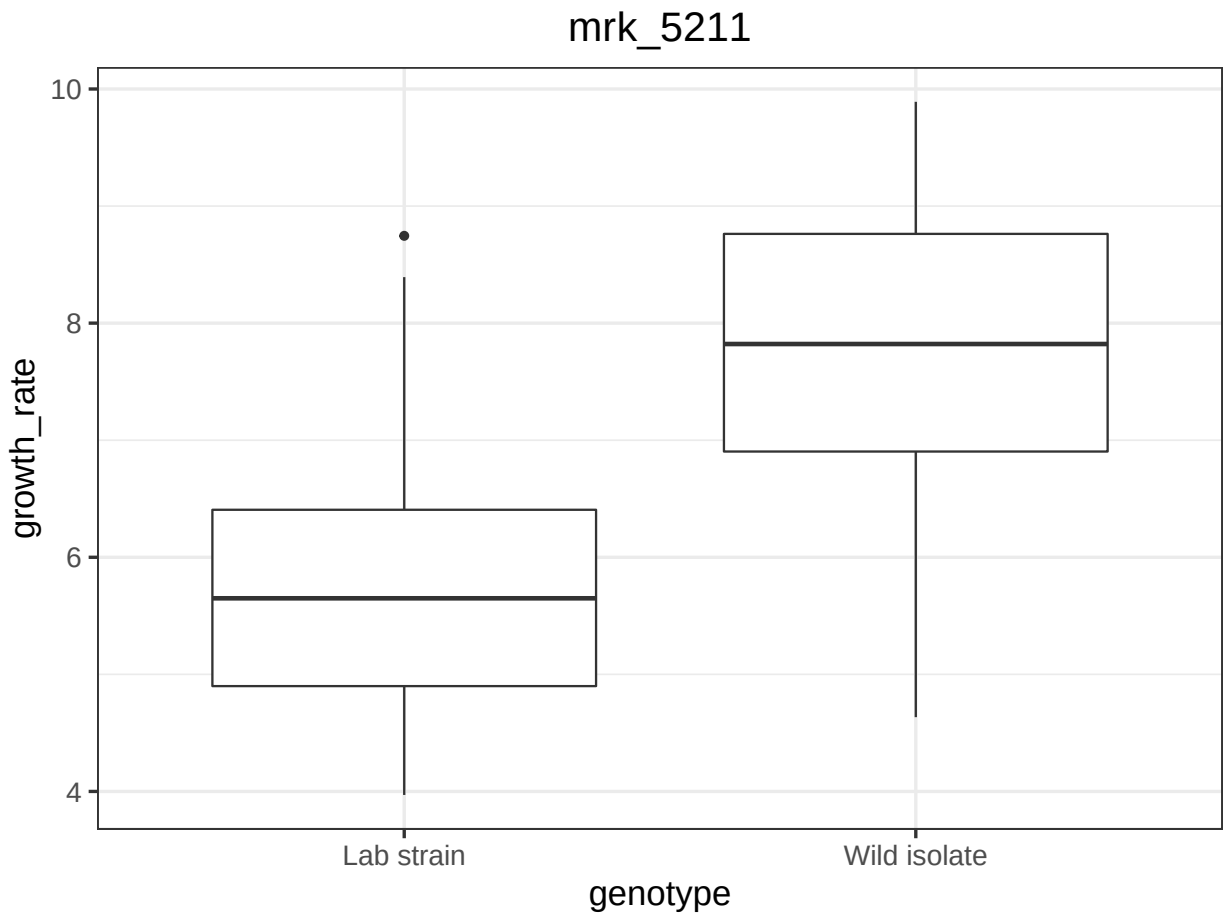
1. Load yeast data and required packages

```
library(ggplot2)
library(data.table)
library(magrittr)
library(tidyr)
library(dplyr)
library(BBmisc)
gene <- fread("./extdata/eqtl/gene.txt")
genotype <- fread("./extdata/eqtl/genotype.txt")
genotype <- melt(genotype, id.vars = 'strain', variable.name = 'marker',
                 value.name = 'genotype')
growth <- fread("./extdata/eqtl/growth.txt")
growth <- melt(growth, id.vars = "strain", variable.name = 'media',
               value.name = 'growth_rate')
marker <- fread("./extdata/eqtl/marker.txt")
```

2. We reproduce the growth boxplot for each marker

```
getMaltoseDt = function(mrk){
  growth_mrk <-  merge(growth, genotype[marker == mrk, .(strain, genotype)],
                       by = 'strain')
  growth_mrk[media == "YPMalt"]
}

# boxplot
plot_growth_one_mk <- function(mk){
    ggplot(getMaltoseDt(mk), aes(genotype, growth_rate)) +
    geom_boxplot() +
    labs(title = mk) + theme_bw(base_size = 16) +
    theme(plot.title = element_text(hjust = 0.5))
}
plot_growth_one_mk("mrk_5211")
```



Section 01 - Test the association between markers and growth

1. Testing marker 5211

Last week, using permutation, we saw that some markers associated with growth. Which of the statistical tests from the lecture would you use to test this association? For each marker, we are not certain if the genotype will cause a positive or negative effect on growth; therefore, which kind of alternative hypothesis would you choose: double-sided, right or left? Apply the test to marker 5211 to obtain a p-value and see if the association that we found last week still holds. If more than one of the tests are appropriate, use them all and compare the results.

2. Applying the test to other markers

Make a function that given a marker and the name of a statistical test, returns the p-value of the association of that marker wrt growth. Test your function on the markers 1653 and 5091. Since we used the same marker last week: do you get similar results as last week?

Section 02 - Pitfalls when using the Student's t-Test

1. Load the data from the **stats-pitfalls.csv** file and visualize the data. Apply both the t-test and Wilcoxon test on it. What do you observe? Which test is the better choice here?

Hint: use stat_summary() or geom_vline() to add points/lines to the ggplot object.

Section 03 - Test the association between 2 markers

1. Last week, using permutation, we saw that marker 5091 is significantly associated with marker 5211. Which of the statistical tests from the lecture would you use to test this association? Apply it to obtain a p-value and see if the association still holds. Note: association can mean that 2 markers choose the same genotype more often than usual, or that 2 markers choose different genotype more often than usual.

Use the contingency table:

```
mks_geno <- genotype[marker %in% c('mrk_5091', 'mrk_5211')] %>%
  spread(marker, genotype)
table(mks_geno[, 2:3])
```

2. Make a function that given any two markers, returns the p-value of the appropriate statistical test to evaluate the association between the markers. Give the alternative as a parameter of the function. Test your function for, e.g., mrk_1 vs mrk_13314.

Section 04 - Test the association between markers and genotype

1. We know that each marker is composed from 2 different genotypes: lab and wild strains. We assumed that the ratio between them is 1:1 (or 50%-50%). Compute the actual ratio.

2. We are interested in testing whether a marker is associated with a genotype, i.e., if a marker has more lab strain or wild isolate than the expected 50%. Which of the statistical tests from the lecture would you use to test this association? Apply it to markers 3385 and 13314 to obtain a corresponding p-value. Make a function out of it.

3. So far, we have hand-picked some markers. Apply the corresponding test to all markers and obtain a distribution of p-values. Make a histogram out of that distribution. Further, compute how often we reject the null hypothesis that the ratio is 0.5 at the 5% level. What do you observe?

Section 05 - Correlations and correlation tests

1. Investigate the correlation between `Sepal.Length` and `Sepal.Width` within the iris dataset. Calculate the correlation and plot your results. Are there any issues with your results? Discuss.

2. Repeat the previous analysis for each species independently. How does the correlation between `Sepal.Length` and `Sepal.Width` change?

3. We want to know whether there is a correlation between attendance to the exercise sessions and the points achieved in the final exame of the students in our course based on the previous years' data. We provide simulated data below (due to legal restrictions). Load the data from **exam_correlation.tsv**. Calculate the correlation between attendance and points using `Pearson` and `Spearman` methods and visualize it. Some students will drop out of the distribution since they were planning to take the retake exam and skipped the first exam, thus obtaining a grade of zero. Which correlation method should be preferred in this context and why?